PISM POLSKI INSTYTUT SPRAW MIĘDZYNARODOWYCH
THE POLISH INSTITUTE OF INTERNATIONAL AFFAIRS

NO. 63 (2564), 18 JUNE 2025 © PISM

BULLETIN

Disinformation Actors Already Targeting AI

Aleksandra Wójtowicz

Artificial intelligence (AI) is already widely used by spreaders of disinformation. While initially the greatest concern was the use of deepfakes, a growing threat is the manipulation of generative AI models, including by tampering with the datasets on which they are trained to get them to publish false or misleading content. Key to building resilience is securing datasets, strengthening moderation, and adapting EU and national regulations to new challenges.

AI and Disinformation. Generative AI, the kind that creates new content, supports influence operations by automating content production. Language models in turn are used to generate seemingly coherent narratives, false or misleading articles, comments, and real-time responses on social media. Disinformation actors are also increasingly using AI to coordinate networks of troll and bot accounts (accounts through they employ mass, synchronised distribution of material) for faster creation and easier management of their content in real time.

While initially deepfake material (AI-generated or manipulated images, audio, or video content) was of great concern, its relevance in operational practice remains limited today. Producing realistic audio or video recordings is expensive and more difficult to disseminate than generating textual narratives for social media. Research (Dabrowska, 2024) indicates that deepfakes most often appear as memes in election campaigns. Instead of videos that mimic reality well, a kind of deepfake called "cheapfakes" because of its low quality are more often used. However, this does not mean that the threat has been eliminated because the technology is rapidly developing and there is great potential for use in influence operations (e.g., creating credible recordings imitating candidates to sway elections). Deepfake technology is also used to create personas, or false identities for troll accounts, that are intended to appear authentic.

The practical application of these technologies can be seen in the ongoing Russian operations *Doppelganger* and *Overload*, which use AI to mass produce content that promoted specific narratives, often anti-Western or pro-Russian. *Doppelganger*, considered one of the largest disinformation operations against the West, involves the creation of copycat websites and profiles on social media, mimicking known, credible Western media. The manipulated content published on them though, is intended, among other things, to undermine support for Ukraine, create distrust of Western institutions, and influence electoral processes in EU and NATO countries. *Overload* involves the mass sending of false or manipulated content such as emails, posts, or videos to editors, fact-checkers, and researchers, often using copied logos and AI-generated expert voices of legitimate sources. Its aim was to *overload* and create confusion in media environments. *Overload* outlets impersonate the editorial sections of media such as EuroNews, Deutsche Welle, and RMF FM.

Disinformation to Train AI. The use of disinformation to deliberately contaminate the datasets on which AI models are trained is becoming an increasingly serious threat. Due to the inability to fully validate huge datasets, AI can also learn from manipulated or false content. As early as 2023, a study by Stanford's Internet Observatory showed that AI training data is sometimes contaminated with, among other things, child pornography. Nowadays, disinformation actors are trying to tamper with the content of the collections both by creating their own databases and by contaminating existing ones.

Russia actively manipulates the collection by deliberately introducing disinformation and false narratives. A key tool of this strategy is the extensive *Pravda* network, which as of 2022 publishes millions of articles on hundreds of websites posing as local news portals in dozens of countries. This

content, often copied from Russian state media and translated automatically, is massively distributed to increase the presence of pro-Russian narratives in the global news ecosystem. The scale of the operation means that the content it generates ends up in open datasets used by developers of large-scale language models (LLMs) and other AI systems. As a result, the model ceases to recognise disinformation—AI encountering many articles confirming a thesis (Russian disinformation, in fact) recognises it as true and begins to replicate and amplify it—often quoting the false sources or suggesting they are credible. NewsGuard's 2025 research found that more than 33% of the chatbots it tested repeated pro-Russian disinformation and 70% cited fabricated articles.

Generative models also learn through user interactions. Disinformation actors game this system to provide manipulated narratives that can influence the model despite moderation. It involves indiscriminate so-called prompt-injection attacks involving misleading information via queries and searches in generative AI models (e.g., Chat GPT), which can lead to subtle distortions in the generated content that are difficult to detect.

Response. In light of the growing threats from disinformation and AI, the EU adopted the Digital Services Act (DSA) and the Artificial Intelligence Act (AI Act). While these are primarily intended to increase the transparency of algorithms and the security of users, they do not fully address the potential of AI to manipulate or present disinformation content. The AI Act classifies AI systems according to potential risk. In the case of generative models, which are designed to interact with humans or create images or sounds (those that can facilitate impersonation and manipulation), a clear indication that the content in question was created using AI is required. In contrast, in the DSA, artificial intelligence has been recognised as a systemic risk, but the regulations do not focus directly on AI. Regulatory response is key, not least to set the tone for specific usage policies put in place by companies.

Entities such as OpenAI, Meta, and Google have changed their usage policies, restricting the use of their products for disinformation purposes. OpenAI banned the impersonation of candidates in elections using their tools and updated their security policies, although paradoxically they stopped considering disinformation as a critical threat. Meta introduced labelling of AI-generated content in May 2024, but this mechanism is mainly based on voluntary labelling by Facebook or Instagram users. At the same time, teams of specialists for countering influence operations are being set up and content moderation mechanisms are being expanded.

Conclusions and Recommendations. Artificial intelligence is not only changing the way disinformation operations are conducted, but is itself becoming their target. Modern language models, despite their enormous potential, are vulnerable to manipulation, which can undermine their credibility and exacerbate information chaos. The response to these challenges must be comprehensive, combining regulatory change, technological action, user education, and close cross-sectoral cooperation. However, what is still lacking is effective collaboration with platforms as well as mechanisms to respond quickly to disinformation incidents involving AI. Moreover, the recent change in the approach of social media platforms to content moderation in the name of protecting what they broadly define as "freedom of speech" is making it more difficult to combat disinformation involving AI.

For the EU, it is crucial to align the DSA with the risks posed by the influence of disinformation actors on generative models. A clear definition of the obligations of digital platforms towards disinformation content created with AI is needed. Although OpenAI as a very large online platform (VLOP) is already subject to the DSA, it is necessary to clarify in the regulation the obligations regarding the moderation of content in AI models.

Equally important is the intensification of dialogue between digital platforms, states and EU institutions to develop rapid pathways to respond to AI-related disinformation and conduct coordinated operations to limit its spread. Examples of such operations include platforms setting up election operations centres, working with fact-checkers, and implementing mechanisms to flag and remove AI-generated content that may mislead users or threaten democratic processes. So far, these have been led by social media platforms. To a greater extent, they should also be implemented in the context of generative models by companies developing AI.

The control of AI model training data through a dual process of database verification needs to be strengthened. It should include audits of the quality and provenance of data inputs conducted by external companies on behalf of platforms, as well as the implementation of AI tools trained on narrower and specialised datasets, capable of automatically detecting false or manipulated content in training sets. It is essential that such control is required by EU law on a similar basis to content moderation in the DSA. Platforms would therefore have to produce the appropriate mechanisms and, if not implemented, could face penalties from the European Commission.